SIMGAN: PHOTO-REALISTIC SEMANTIC IMAGE MANIPULATION USING GENERATIVE ADVERSARIAL NETWORKS

Simiao Yu[⋆] *Hao Dong*[⋆] *Felix Liang*[†] *Yuanhan Mo*[⋆] *Chao Wu*[‡] *Yike Guo*[⋆]

[⋆] Imperial College London
[†]University of Washington
[‡]Zhejiang University

ABSTRACT

Semantic image manipulation (SIM) aims to generate realistic images from an input source image and a target text description, such that the generated images not only match the content of the description, but also maintain text-irrelevant features of the source image. It requires to learn a good mapping between visual features and linguistic features. Previous works on SIM can only generate images of limited resolution that typically lack of fine and clear details. In this work, we aim to generate high-resolution photo-realistic images for SIM. Specifically, we propose SIMGAN, a generative adversarial networks (GAN) based architecture that is capable of generating images of size 256 $\times$ 256 for SIM. We demonstrate the effectiveness of SIMGAN and its superiority over existing methods via qualitative and quantitative evaluation on Caltech-200 and Oxford-102 datasets.

Index Terms— adversarial learning, generative model, image generation, semantic image manipulation

1. INTRODUCTION

Humans have the ability to manipulate representations of imaginary pictures in their minds in a goal-oriented fashion [1]. Such capability is of great significance in generating creative thoughts for visual arts [2]. A natural question is whether machines can have such ability. This can be formulated into an image generation task called semantic image manipulation (SIM) [3]. Specifically, given an input source image and a target text description, the generated images for SIM should 1) match what the target description specifies; 2) maintain text-irrelevant features of the source image, and 3) be realistic and plausible. A key challenge of SIM is to learn a good mapping between visual and linguistic features.

Recently, a number of Generative Adversarial Networks (GAN) [4] based generative models [5, 3, 6, 7] have been proposed to tackle SIM. They have succeeded in generating images of size up to 128 $\times$ 128 for SIM. However, the generated images typically lack of fine and clear details due to their relatively low resolution. Generating images of higher resolution with photo-realistic details, and with aforementioned requirements of SIM satisfied, remains a challenge.

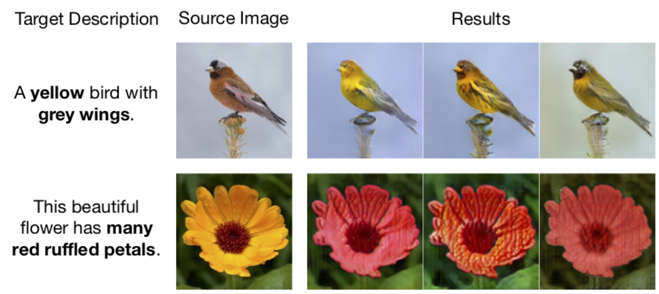

Fig. 1. Photo-realistic images for semantic image manipulation (SIM) generated by our proposed SIMGAN.

In this work, we therefore aim to generate photo-realistic images for SIM. Specifically, we propose SIMGAN, a novel GAN based framework for SIM that is capable of synthesising high-resolution 256 $\times$ 256 images. The main idea is that in training stage we reuse the same generator (which is originally used to generate images for SIM given a source image and a target text description) to map those generated images back to its corresponding source image with the ground-truth matching text description. Inspired by previous works on neural machine translation [8] and image-to-image translation [9, 10, 11], such introduced cycle-consistent constraint can significantly reduce the required search space of the generating function. As a result, the generator can not only synthesise higher-resolution images (which would be difficult without such imposed constraint, as high-resolution generated images make it easily distinguished from real images [12, 13]), but also better maintain the text-irrelevant contents (e.g. backgrounds) of source images and match target text descriptions, as required by SIM. Some results of SIM generated by our proposed SIMGAN are presented in Fig. 1.

The main contribution of this work is the design of SIMGAN, which is able to generate photo-realistic 256 $\times$ 256 images for SIM. To the best of our knowledge, this is the best result reported yet for this task. We evaluate our model on Caltech-200 [14] and Oxford-102 [15] datasets, and demon-

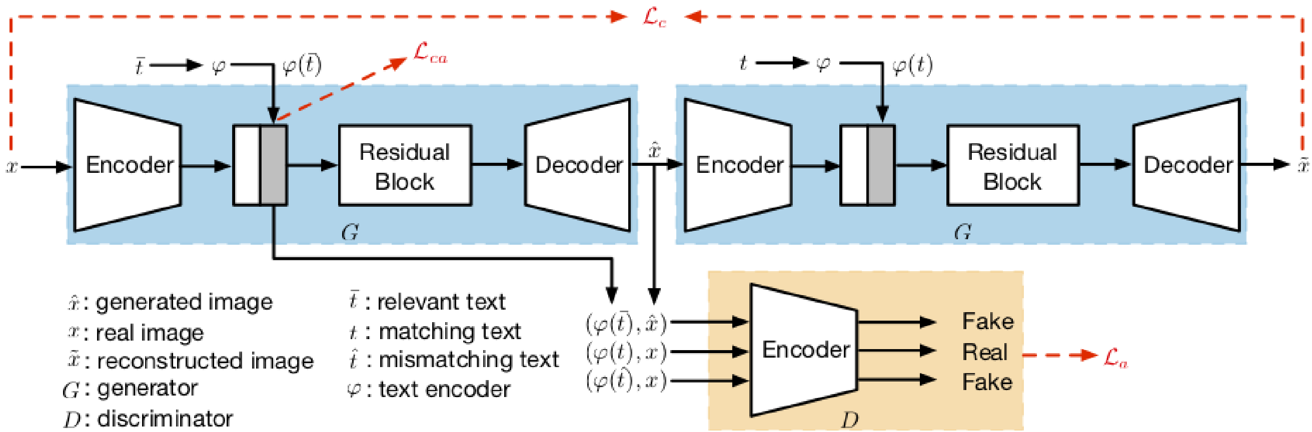

Fig. 2. Schema of our proposed SIMGAN architecture for semantic image manipulation (SIM).

strate the effectiveness of SIMGAN and its superiority over other comparison methods, in terms of generating sharp and realistic images that contain clear and fine details and matching the semantics specified by target text descriptions.

1.1. Prior work

Reed *et al.* [5] first proposed a two-step method, in which an auxiliary style encoder was used to invert the trained generator of text-to-image synthesis, so that the text-irrelevant features of source images can then be extracted. Our previous work [3] employed the cGAN framework [16] directly conditioned on both image and text information, with an adaptive loss developed for SIM. Nam *et al.* [7] proposed to use a text-adaptive discriminator, which consisted of multiple word-level local discriminators that can disentangle fine-grained visual attributes from text descriptions. Similar to this work, Liu *et al.* [6] recently proposed to impose cycle consistency for SIM. However, compared with their approach, our proposed SIMGAN has three main advantages. First, SIMGAN is able to generate images of much higher resolution (256×256 *vs.* 64×64), with much finer and clearer photo-realistic details. Second, SIMGAN enables one-to-many generation, i.e. multiple and diverse images can be generated given one source image and one target description. Last, SIMGAN only needs to train one set of generator and discriminator, which leads to a much more compact model.

2. METHOD

The schema of SIMGAN architecture is illustrated in Fig. 2. It consists of a generator G and a discriminator D. The generator G has three modules: an encoder, a residual block and a decoder.

At inference stage, the trained generator G^* will first take as input a source image x (whose features will then be extracted by the encoder module) and features of a target text description $\varphi(\bar{\mathrm{t}})$ (extracted by a pretrained text encoder φ). Extracted features of the image and text will then be concatenated and fed to the residual block. The use of the residual block will not only help retain underlying structure of x as required by SIM, but also enable the model via a deeper encoding process to learn better mappings between visual and textural features [3]. Finally, the output of the residual block will be as the input of the decoder module, from which multiple and diverse images $\hat{\mathrm{x}}$ for SIM will then be generated.

At training stage, several designated loss terms are employed to enable SIMGAN to generate high-resolution photo-realistic images for SIM. These will be explained in detail as follows.

2.1. Adversarial loss

The generated images for SIM are required to not only be realistic and match the given text descriptions, but also maintain the text-irrelevant features of the source images. It would be difficult to explicitly define a corresponding learning objective for such complicated image generation problem. Instead, we employ adversarial learning to implicitly learn an adaptive loss function for SIM. The discriminator D receives three types of inputs: real images x with matching texts t (as real score), real images with mismatching texts $\tilde{\mathrm{t}}$ (as fake score) and generated images $\hat{\mathrm{x}}$ with relevant texts $\bar{\mathrm{t}}$ [3] (as fake score). Also, we employ the least square loss [17] rather than the original negative log likelihood, in order to improve the stability of GAN training and to generate images of better quality and higher resolution. The adversarial loss $\mathcal{L}_a$ of SIMGAN is defined as follows.

$$\begin{aligned}\mathcal{L}_a(G, D) = &\mathbb{E}_{(\mathrm{x},\mathrm{t})\sim p_{data}}[(D(\mathrm{x}, \varphi(\mathrm{t})) - 1)^2] \\ &+ \mathbb{E}_{(\mathrm{x},\tilde{\mathrm{t}})\sim p_{data}}[D(\mathrm{x}, \varphi(\tilde{\mathrm{t}}))^2] \\ &+ \mathbb{E}_{(\mathrm{x},\bar{\mathrm{t}})\sim p_{data}}[D(G(\mathrm{x}, \varphi(\bar{\mathrm{t}})), \varphi(\bar{\mathrm{t}}))^2].\end{aligned} \quad (1)$$

2.2. Cycle loss

Using the adversarial loss alone is not sufficient to generate higher-resolution images for SIM, because generated images of higher-resolution can be distinguished more easily from real images by the discriminator D, which will in turn make the training process unstable [12, 13] and thus impair the quality of generation. To tackle this challenge, inspired by recent works on machine translation [8] and image-to-image generation [9, 10, 11], at training stage we reuse the generator G to reconstruct the input source image $\tilde{\mathrm{x}}$ from the generated images $\hat{\mathrm{x}}$ (with the matching text t of x, see Fig. 2), and apply a cycle loss $\mathcal{L}_c$ to enforce $\tilde{\mathrm{x}}$ to be close to x. We define $\mathcal{L}_c$ as following:

$$\mathcal{L}_c(G) = \mathbb{E}_{(\mathrm{x},\mathrm{t},\bar{\mathrm{t}})\sim p_{data}}[\|G(G(\mathrm{x},\varphi(\bar{\mathrm{t}})),\varphi(\mathrm{t}))-\mathrm{x}\|_1]. \quad (2)$$

This cycle loss significantly reduces the searching space of the generator G, which critically contributes to the generation of photo-realistic images for SIM.

2.3. Conditioning augmentation loss

SIM essentially is a one-to-many generation task. To enable SIMGAN to generate diverse images for SIM given one source image and one target text description, we apply the method of conditioning augmentation [18]. Specifically, it allows additional text features to be sampled from a Gaussian distribution $\mathcal{N}(C_\mu(\varphi(\mathrm{t})), C_\Sigma(\varphi(\mathrm{t})))$, in which its mean and diagonal covariance matrix are functions of target text features $\varphi(\mathrm{t})$ (denoted as C_μ and C_Σ respectively) with learnable parameters trained along with the model. A conditioning augmentation loss $\mathcal{L}_{ca}$ is incorporated in SIMGAN:

$$\mathcal{L}_{ca}(C_\mu, C_\Sigma) = \mathrm{KL}(\mathcal{N}(C_\mu(\varphi(\mathrm{t})), C_\Sigma(\varphi(\mathrm{t}))) \parallel \mathcal{N}(0, I)), \quad (3)$$

which is the Kullback-Leibler (KL) divergence between the standard Gaussian distribution and the target sampling Gaussian distribution.

It is worth noting that we only apply $\mathcal{L}_{ca}$ in the process of generating $\hat{\mathrm{x}}$ (from x and $\bar{\mathrm{t}}$) to encourage diversity. We do not impose $\mathcal{L}_{ca}$ when generating the reconstructed image $\tilde{\mathrm{x}}$, in that the cycle loss $\mathcal{L}_c$ enforces $\tilde{\mathrm{x}}$ to match closely the source image x. In such case, therefore, $\tilde{\mathrm{x}}$ should not be of variety.

2.4. Full objective

The full objective function of our proposed SIMGAN is:

$$\begin{aligned}\min_{G,C_\mu,C_\Sigma}\max_D \mathcal{L}(G,D) = \lambda_a\mathcal{L}_a(G,D)+ \\ \lambda_c\mathcal{L}_c(G) + \lambda_{ca}\mathcal{L}_{ca}(C_\mu, C_\Sigma),\end{aligned} \quad (4)$$

where we employ λ_a, λ_c, and λ_{ca} to control the strength of each individual loss term.

3. EXPERIMENTS

3.1. Experimental details

We evaluated our proposed SIMGAN on Caltech-200 [14] and Oxford-102 [15] datasets. We compared SIMGAN with existing approaches for SIM, including SISGAN [3], CCGAN [6] and TAGAN [7].

In the generator G, the encoder had 3 convolutional layers; the residual block had 16 residual units [19], each of which contained 2 convolutional layers; the decoder had 2 transposed convolutional layers. The discriminator D had 8 convolutional layers. ReLU and leaky-ReLU activation were respectively employed in G and D. Batch normalisation [20] was used in all networks. We adopted a pretrained text encoder φ from [21].

We set $\lambda_a = 1$, $\lambda_c = 10$ and $\lambda_{ca} = 1$ in Equation 4. We used the Adam optimiser [22], with an initial learning rate of 0.0002. The networks were trained for 600 epochs, and the learning rate was halved every 50 epochs. Batch size was set to 16. We implemented our methods using TensorFlow [23] and TensorLayer [24].

3.2. Qualitative results

The qualitative results of SIMGAN and the comparison methods are presented in Fig. 3. Although SISGAN and CCGAN are able to generate images that meet the requirements of SIM to some extent, their relatively low resolution limits the quality of the results, which typically lacks of clear and fine details. TAGAN can generate much clearer images and better preserve the text-irrelevant features than those of SISGAN and CCGAN. By contrast, our proposed SIMGAN is capable of not only generating images of highest resolution with most photo-realistic details, but also well matching the target text descriptions while maintaining the features of source images not specified by texts.

3.3. Quantitative results

For SIM, the most reliable quantitative evaluation is human evaluation, though subjective factors may be involved. We hence performed a human study to compare the methods quantitatively. Specifically, we recruited 13 subjects, each of whom was presented 6 images with 6 different target descriptions and the corresponding generated images by each comparison method. The subjects were then asked to rank these four methods (from 1 to 4, 1 for the best) based on three criteria: whether the generated images 1) are of high quality with fine details (sharpness), 2) match well the target text descriptions (accuracy), and 3) preserve the text-irrelevant features of the source images (consistency).

The averaged ranking scores are shown in Table 1. SIMGAN achieves the best ranking scores of sharpness and accuracy among the comparison methods, which indicates its

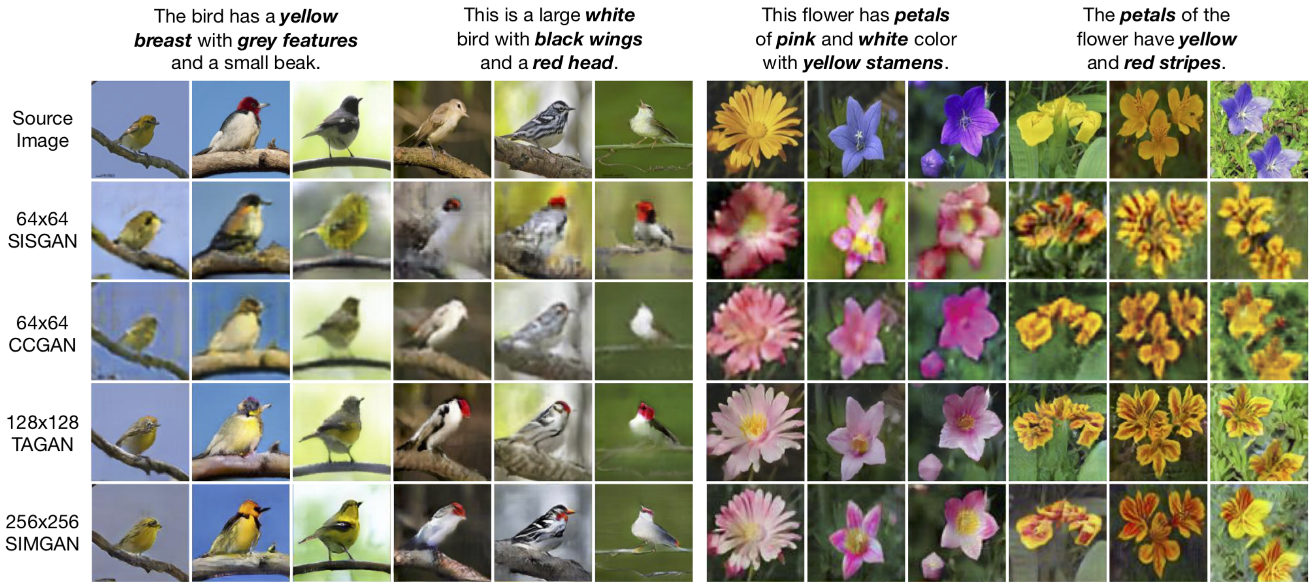

Fig. 3. Qualitative comparison results of SISGAN [3], CCGAN [6], TAGAN [7] and our proposed SIMGAN.

Table 1. Comparison results of human evaluation.

	Sharpness	Accuracy	Consistency
SISGAN [3]	3.32 ± 0.52	2.58 ± 1.00	3.49 ± 0.82
CCGAN [6]	3.45 ± 0.56	3.71 ± 0.58	3.01 ± 0.89
TAGAN [7]	1.68 ± 0.49	2.03 ± 0.66	1.64 ± 0.58
SIMGAN	1.28 ± 0.50	1.71 ± 0.80	1.86 ± 0.77

effectiveness for SIM. However, it has slightly lower score of consistency than that of TAGAN. This could be explained by a fact that the cycle loss used in our model may not well preserve the colour of source images, as also reported in [9].

3.4. Interpolation and diversity results

If traversing in the learned latent space leads to semantic changes to the generated images (i.e. a smooth latent space), then we can reason that the model successfully learns relevant and useful representations, rather than simply memorising the training data [25]. We fed the linearly interpolated representations of two different target text descriptions with a same source image to SIMGAN. As presented in Fig. 4, the generated images clearly demonstrate smooth variations between semantics of the two target descriptions. Moreover, all the generated images not only remain plausible with photo-realistic details, but also maintain the text-irrelevant features of the source image. Such continuous transition indicates that SIMGAN learns relevant and useful features for SIM.

Due to the applied conditioning augmentation loss $\mathcal{L}_{ac}$,

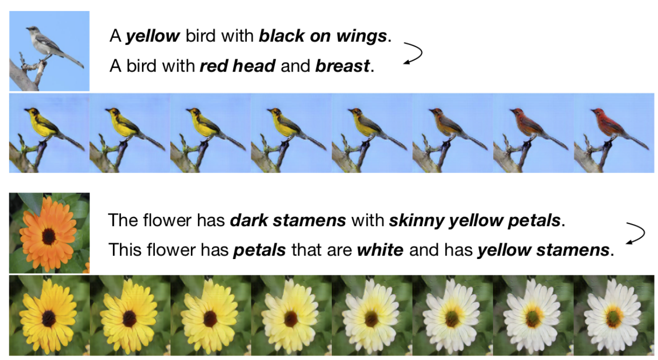

Fig. 4. Results of linearly interpolating between features of two target text descriptions from SIMGAN.

SIMGAN is able to generate diverse images from one source image and target description for SIM (i.e. one-to-many generation, see Fig. 1)

4. CONCLUSION

In this work, we have proposed SIMGAN, which is able to generate photo-realistic images of size 256×256 for semantic image manipulation. This is for the first time that such high-resolution images can be generated for this challenging image generation task. We have demonstrated the effectiveness of our model and its superiority over other existing methods. In future, we aim to apply our model to other datasets that contain more complicated objects and backgrounds.

5. REFERENCES

[1] Alexander T. Sack and Teresa Schuhmann, "Hemispheric differences within the fronto-parietal network dynamics underlying spatial imagery," *Front Psychol.*, vol. 3, pp. 214, 2012.

[2] Alexander Schlegel, Peter J. Kohler, Sergey V. Fogelson, Prescott Alexander, Dedeepya Konuthula, and Peter Ulric Tse, "Network structure and dynamics of the mental workspace," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 40, pp. 16277–16282, 2013.

[3] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo, "Semantic image synthesis via adversarial learning," in *ICCV*, 2017.

[4] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[5] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," in *ICML*, 2016.

[6] Xiyan Liu, Gaofeng Meng, Shiming Xiang, and Chunhong Pan, "Semantic image synthesis via conditional cycle-generative adversarial networks," in *ICPR*, 2018.

[7] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim, "Text-Adaptive generative adversarial networks: manipulating images with natural language.," in *NeurIPS*, 2018.

[8] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma, "Dual learning for machine translation," in *NIPS*, 2016.

[9] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[10] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *ICML*, 2017.

[11] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong, "DualGAN: unsupervised dual learning for image-to-image translation," in *ICCV*, 2017.

[12] Augustus Odena, Christopher Olah, and Jonathon Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *ICML*, 2017.

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *ICLR*, 2018.

[14] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," Tech. Rep. CNS-TR-2011-001, California Institute of Technology, 2011.

[15] Maria-Elena Nilsback and Andrew Zisserman, "Automated flower classification over a large number of classes," in *ICCVGIP*, 2008.

[16] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[17] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017.

[18] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas, "StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[20] Sergey Ioffe and Christian Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.

[21] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in *TACL*, 2015.

[22] Diederik Kingma and Jimmy Ba, "Adam: a method for stochastic optimization," in *ICLR*, 2014.

[23] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, "TensorFlow: a system for large-scale machine learning," in *OSDI*, 2016.

[24] Hao Dong, Akara Supratak, Luo Mai, Fangde Liu, Axel Oehmichen, Simiao Yu, and Yike Guo, "TensorLayer: a versatile library for efficient deep learning development," in *ACM-MM*, 2017.

[25] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative Adversarial Networks," in *ICLR*, 2016.